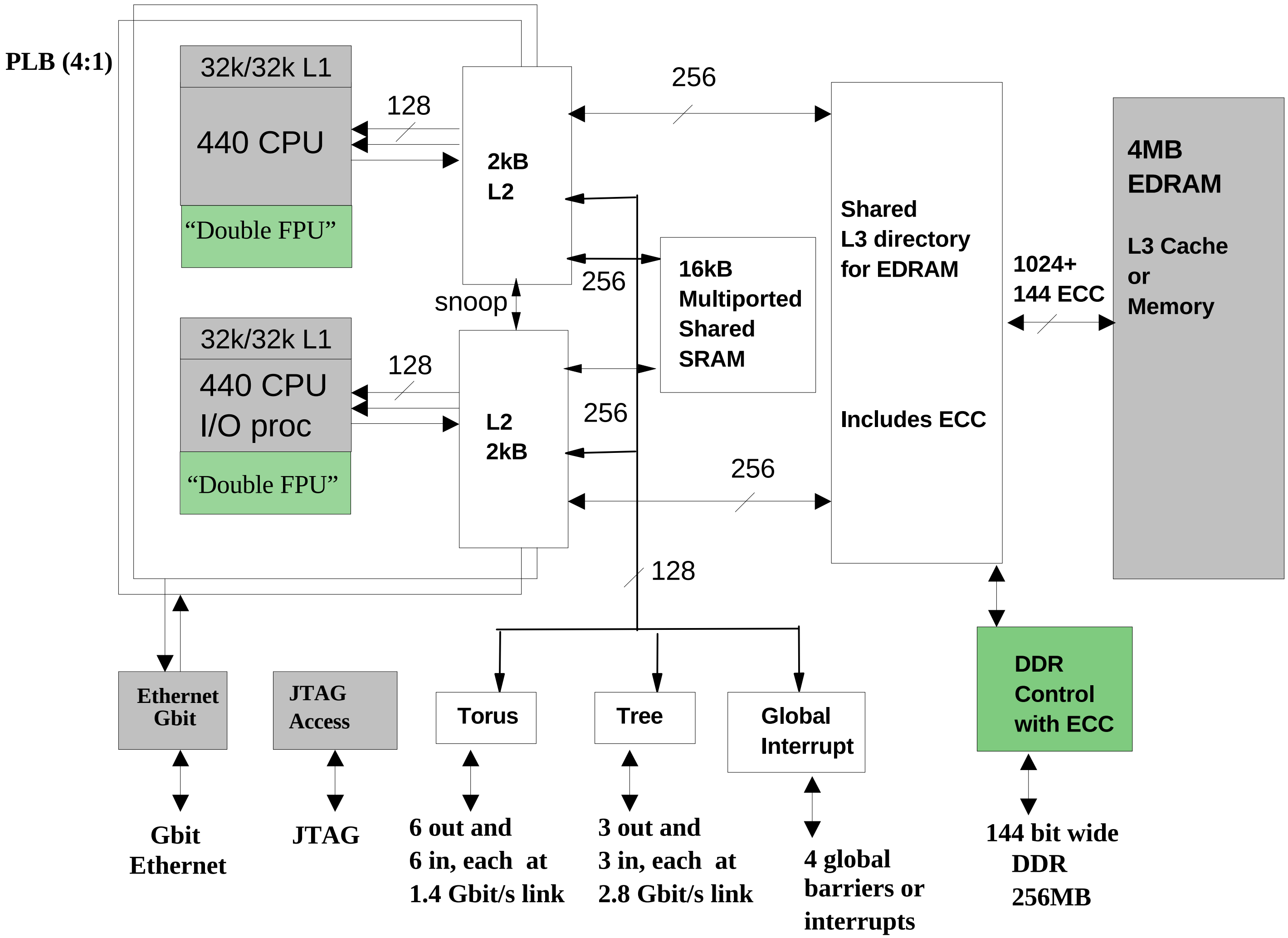
# Microarchitecture and Networks

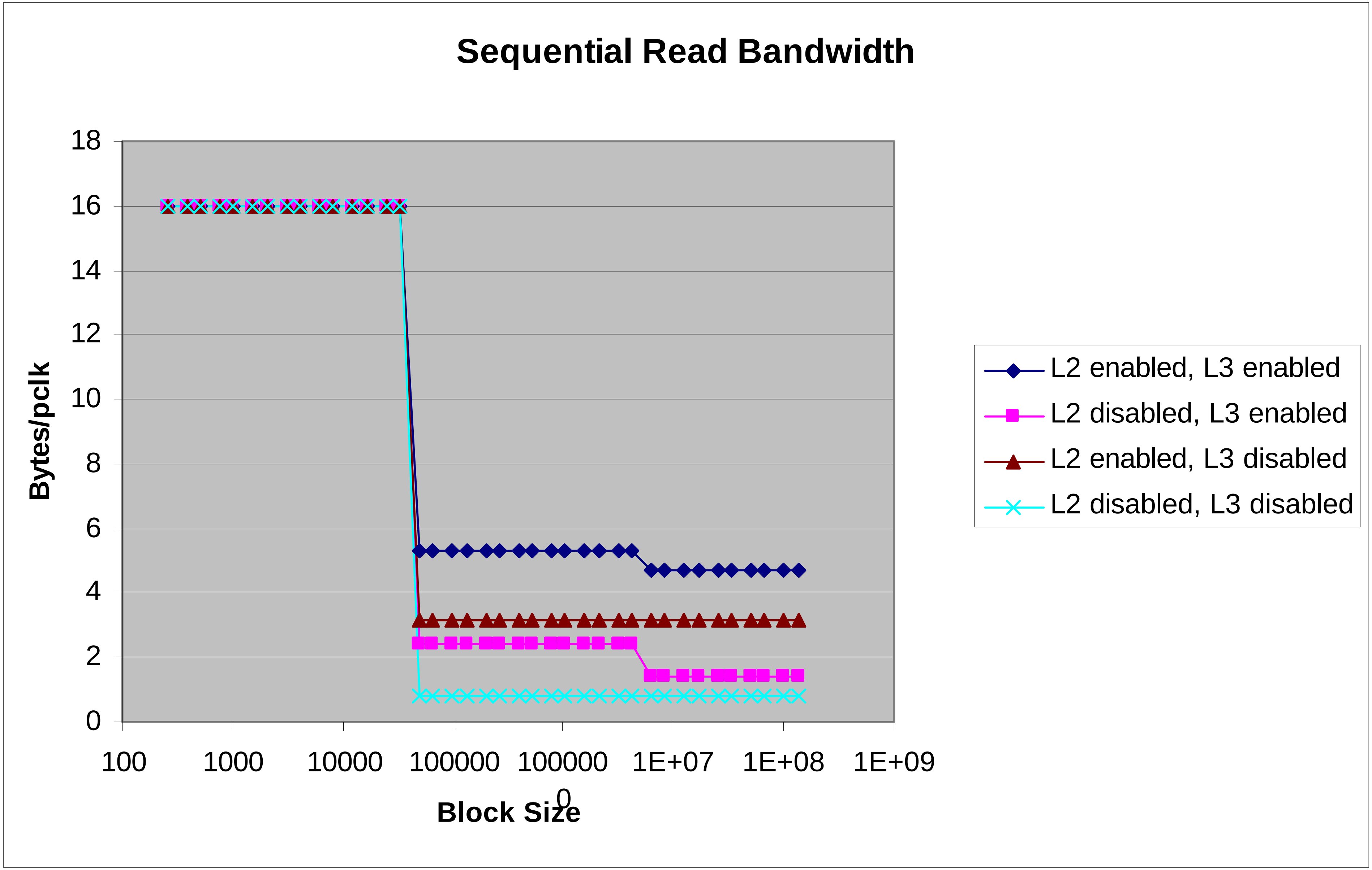
presented by Dirk Hönicke

October 2003 | BlueGene/L Workshop, Reno

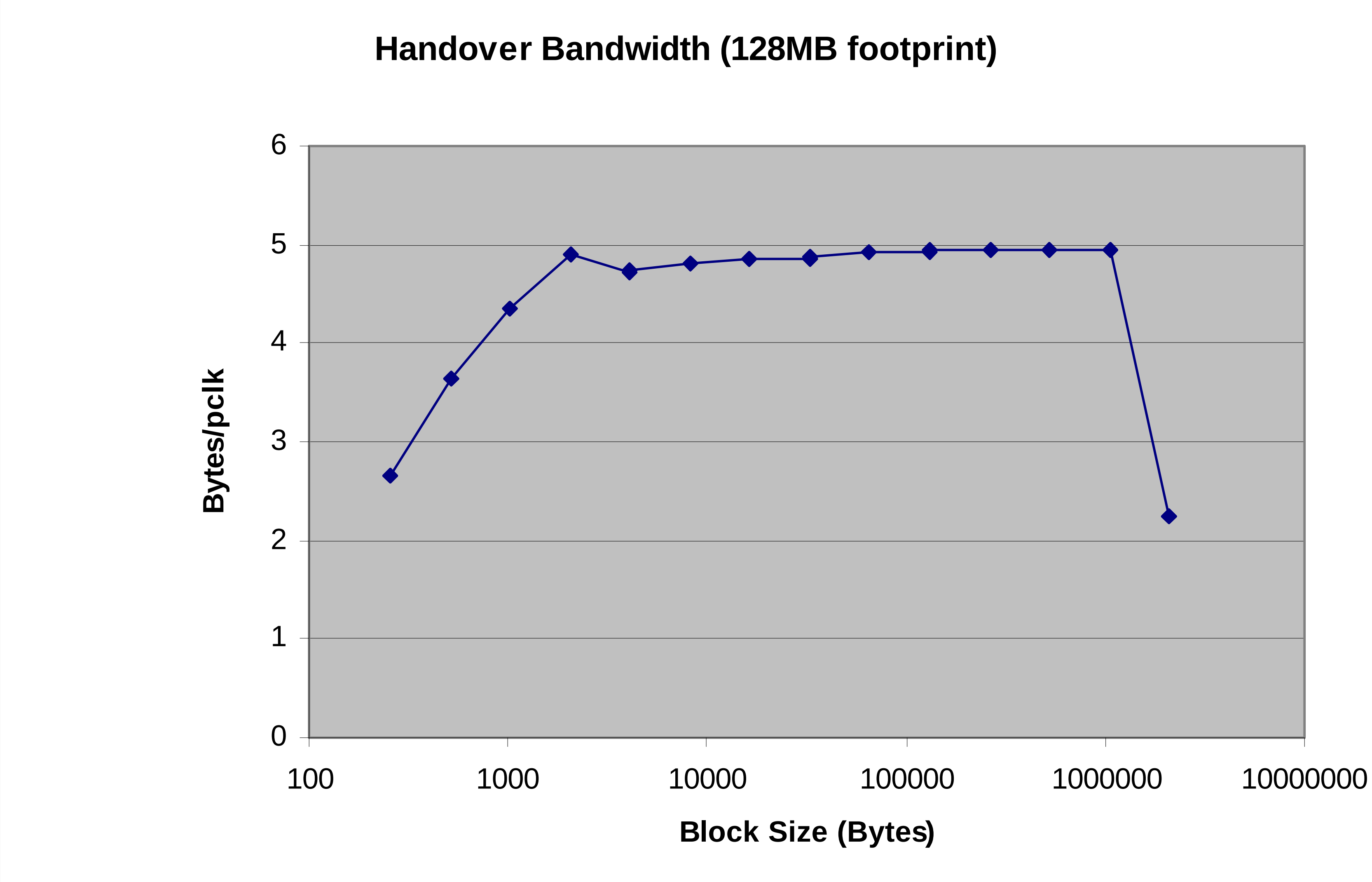

PLB (4:1)
32k/32k L1
440 CPU
"Double FPU"
128
2kB
L2
256
snoop
256
32k/32k L1
440 CPU
I/O proc
"Double FPU"
128
L2
2kB
256
16kB
Multiported
Shared
SRAM
256
Shared
L3 directory
for EDRAM
Includes ECC
1024+
144 ECC
4MB
EDRAM
L3 Cache
or
Memory
128
Ethernet
Gbit
JTAG
Access
Torus
Tree
Global
Interrupt
DDR
Control
with ECC
Gbit
Ethernet
JTAG
6 out and
6 in, each at
1.4 Gbit/s link
3 out and
3 in, each at
2.8 Gbit/s link
4 global
barriers or
interrupts
144 bit wide
DDR
256MB

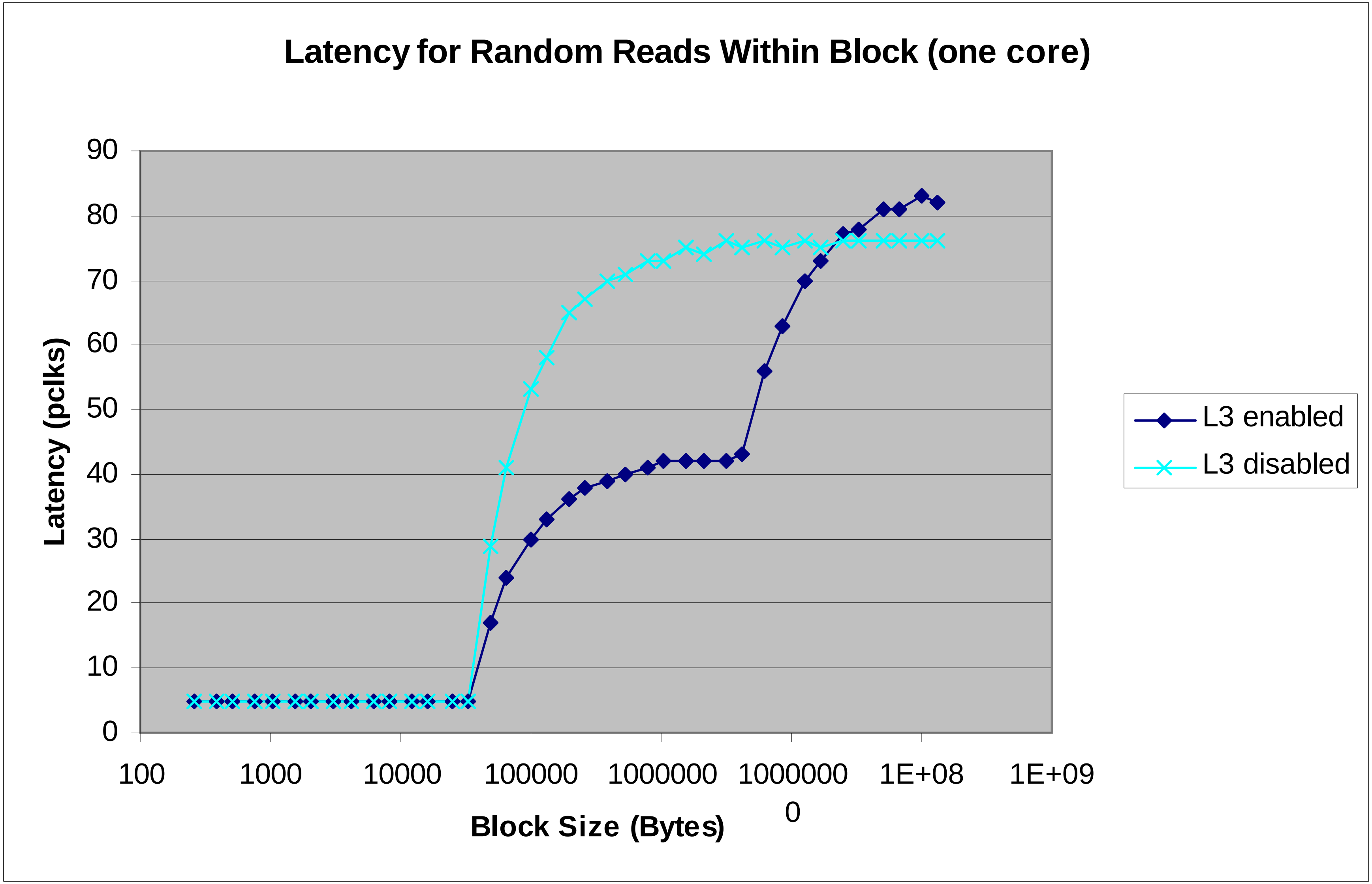
# Some Features

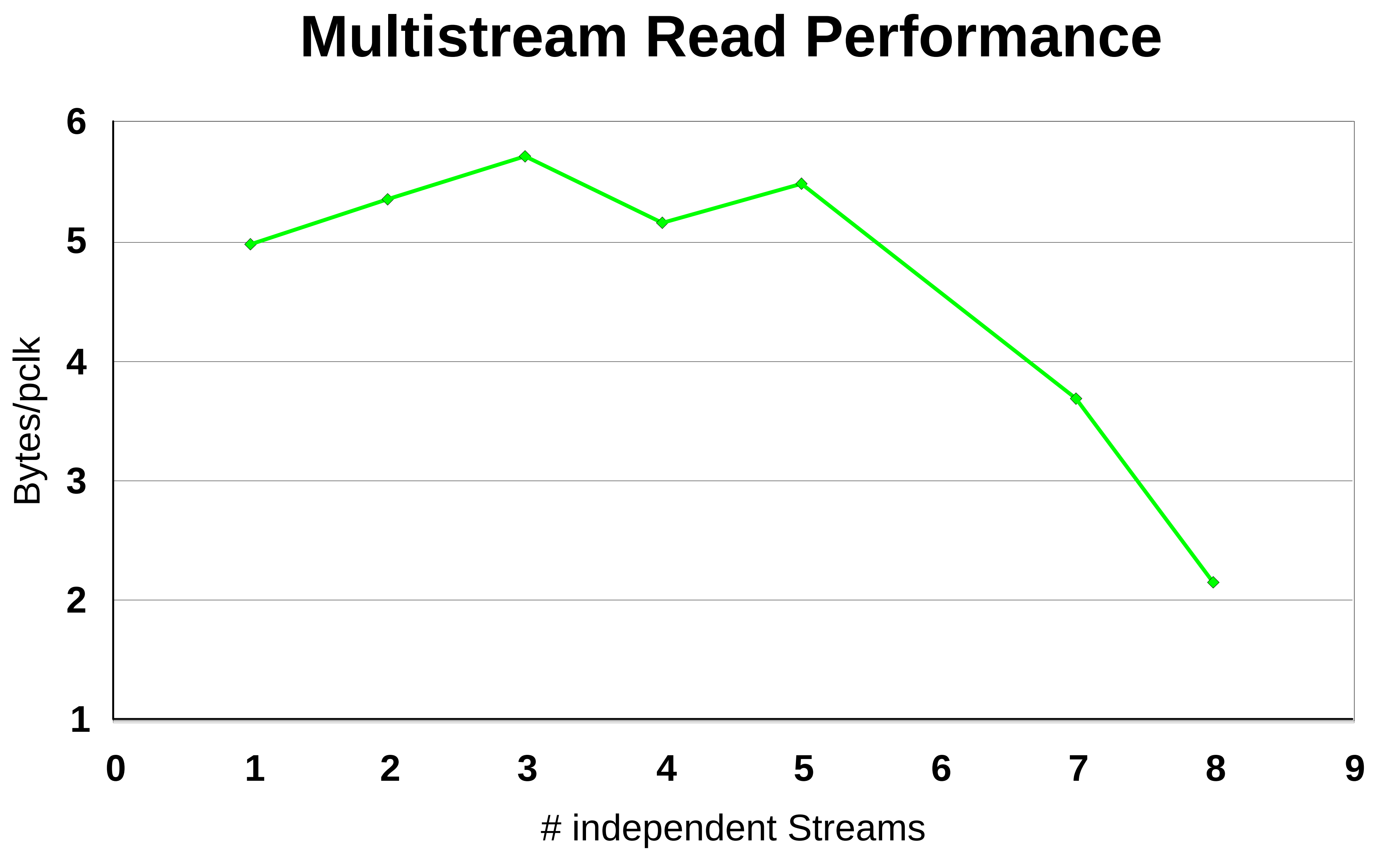
- no MESI, but
- hardware support for
  - flushing L1 cache
  - locks/barriers
  - fast core-to-core signaling (interrupt)
- 16kB SRAM + L3 scratchpad
- fine control over memory access behavior on all memory hierarchy levels

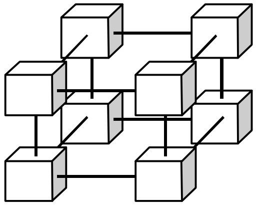

Sequential Read Bandwidth
Bytes/pclk
18
16
14
12
10
8
6
4
2
0
100
1000
10000
100000
100000
0
1E+07
1E+08
1E+09
Block Size
L2 enabled, L3 enabled
L2 disabled, L3 enabled
L2 enabled, L3 disabled
L2 disabled, L3 disabled

## Delivered Memory Bandwidth Between CPUs Sustains Maximum Link Bandwidth



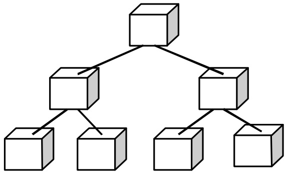

Latency for Random Reads Within Block (one core)
Latency (pclks)
0
10
20
30
40
50
60
70
80
90
100
1000
10000
100000
1000000
1000000
0
1E+08
1E+09
Block Size (Bytes)
L3 enabled
L3 disabled


Multistream Read Performance
Bytes/pclk
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
# independent Streams



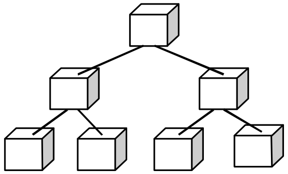
## 3 Dimensional Torus

- Point-to-point



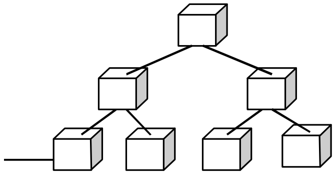
## Global Tree

- Global Operations



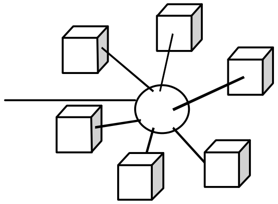
## Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



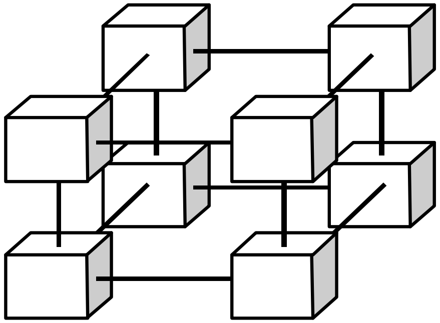
## Gbit Ethernet

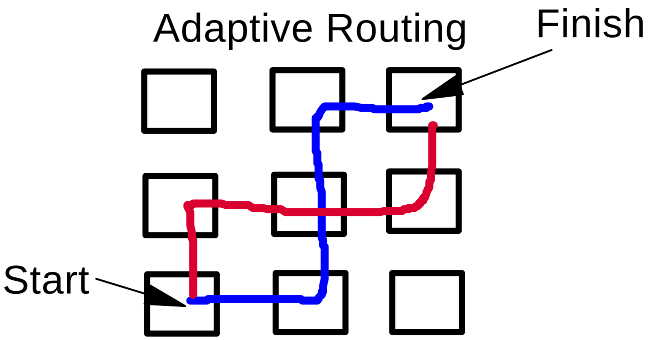
- File I/O and Host Interface



## Control Network

- Boot, Monitoring and Diagnostics





- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * 6 * 1.4Gb/s = 68 TB/s total torus bandwidth
- 4 * 32 *32 * 1.4Gb/s = 5.6 Tb/s Bisectional Bandwidth
- Worst case hardware latency through node ~ 69nsec
- Virtual cut-through routing with multipacket buffering on collision
  - Minimal
  - Adaptive
  - Deadlock Free
- Class Routing Capability (Deadlock-free Hardware Multicast)
  - Packets can be deposited along route to specified destination.
  - Allows for efficient one to many in some instances
- Active messages allows for fast transposes as required in FFTs.
- Independent on-chip network interfaces enable concurrent access.

# Link Utilization on FFT Communication Pattern is above 97%

## All-to-All Efficiency for Mesh & Torus Topologies for BlueGene/L

All messages are 100 packets, each packet 256B

| | 32 (4x4x2) | 512(8x8x8) |
|---|---|---|
| Mesh Time (Processor cycles) | 3.9M | 149.8M |
| % Peak | 88% | 74% |
| Torus Time (Processor cycles) | 1.8M | 56.2M |
| % Peak | 96.8% | **98.3%** |
| Torus Theoretical Bound without idle packets | 1.7M | 55.3M |

# Broadcast in a 2-D plane achieves 98.8% of Peak (4x4 plane)

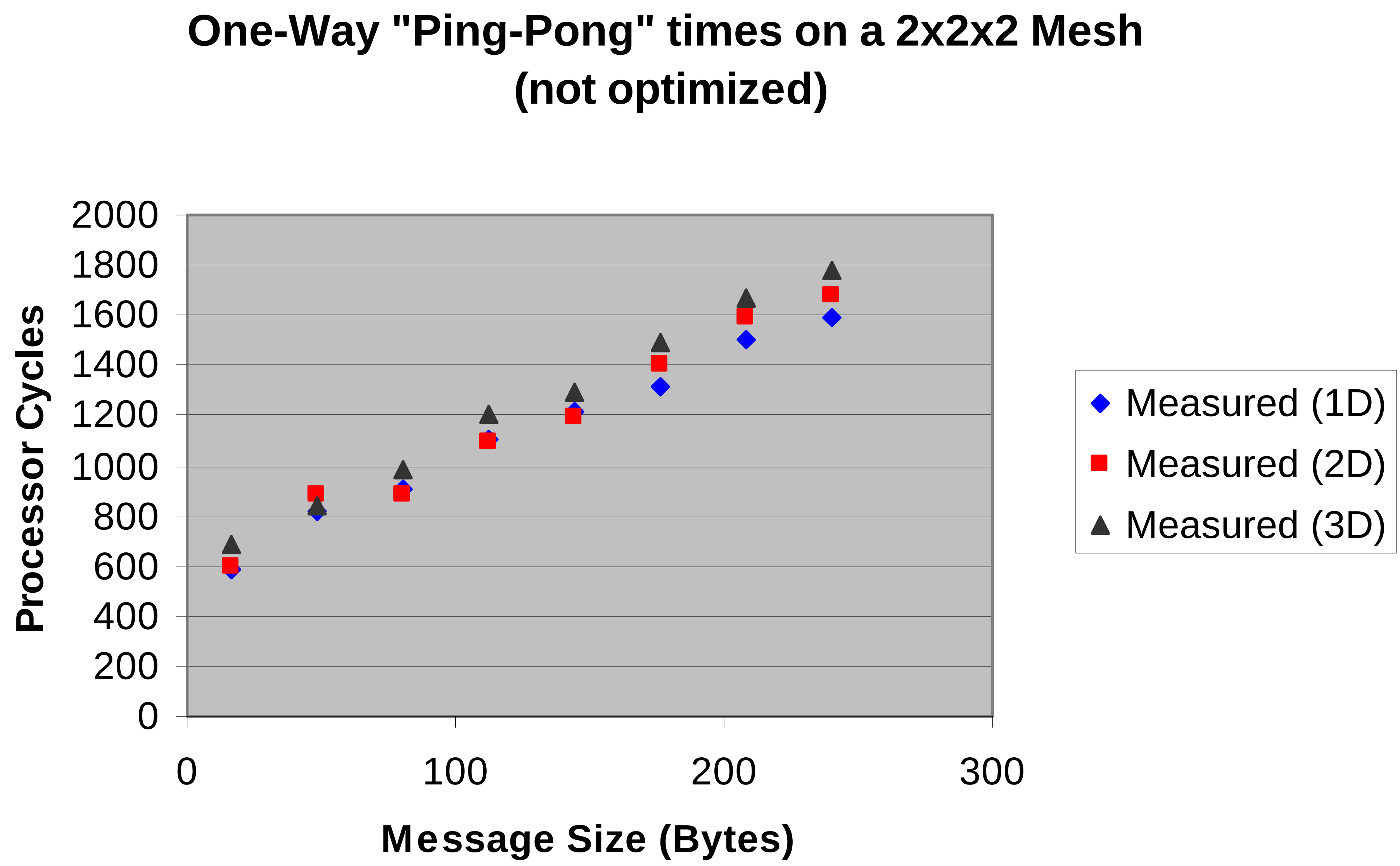
- Torus Network supports Hardware Multicast

  Fundamental to dense solvers

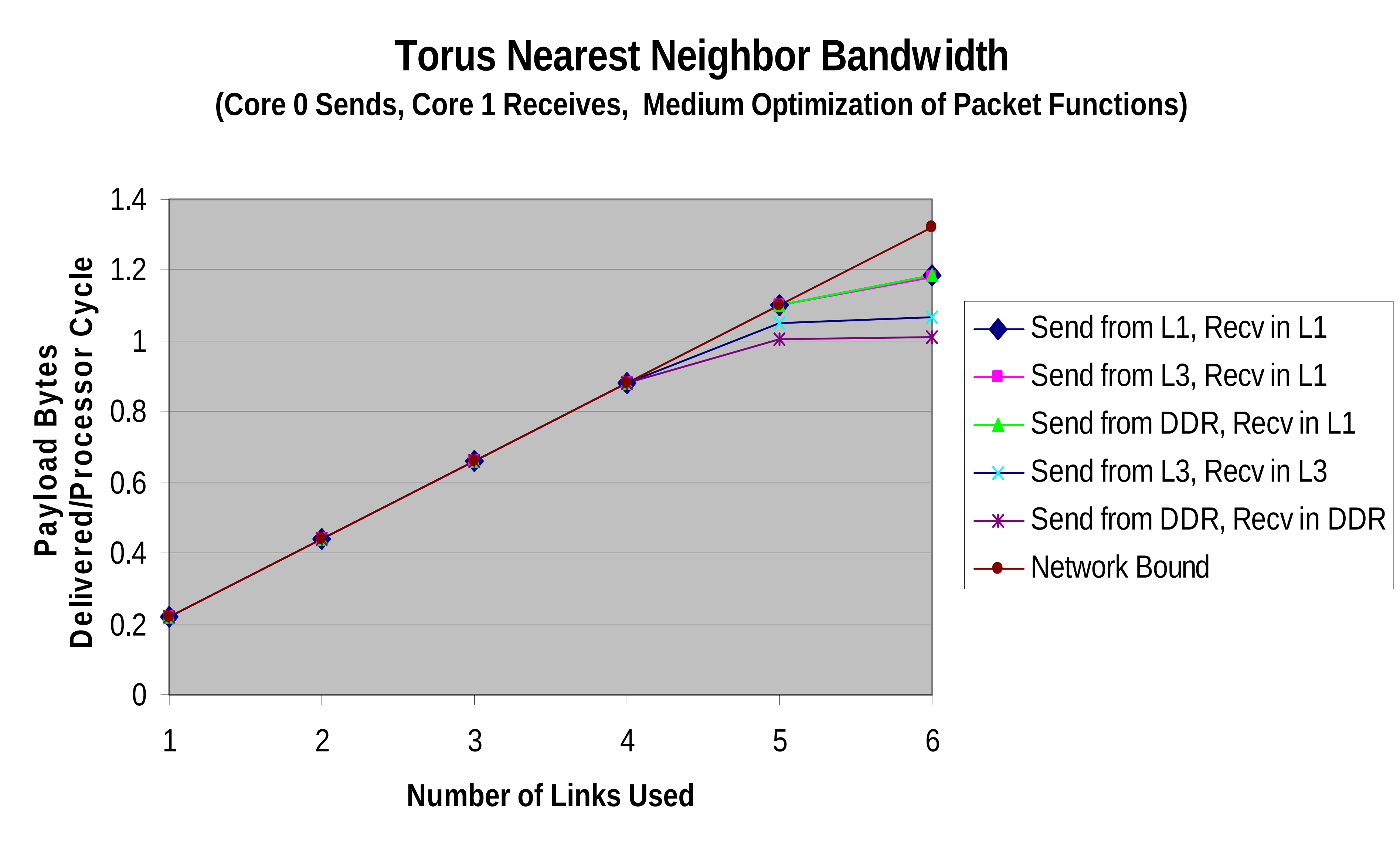
  Confirms network efficiency used in Linpack benchmark efficiency model

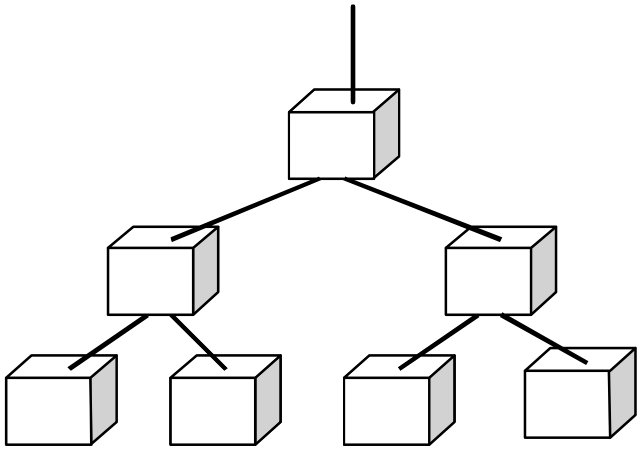
| **Number of Packets** | **Payload (MBytes)** | **Measured (pclks)** | **Theoretcial (pclks)** | **Effciency** |
|---|---|---|---|---|
| 10,000 | 2.4 | 2,727,674 | 2,700,000 | 98.9% |
| 100,000 | 24 | 27,234,320 | 27,000,000 | 99.1% |

# Prototype Delivers ~1usec Ping Pong low-level messaging latency



# Nearest neighbor communication achieves 75-80% of peak



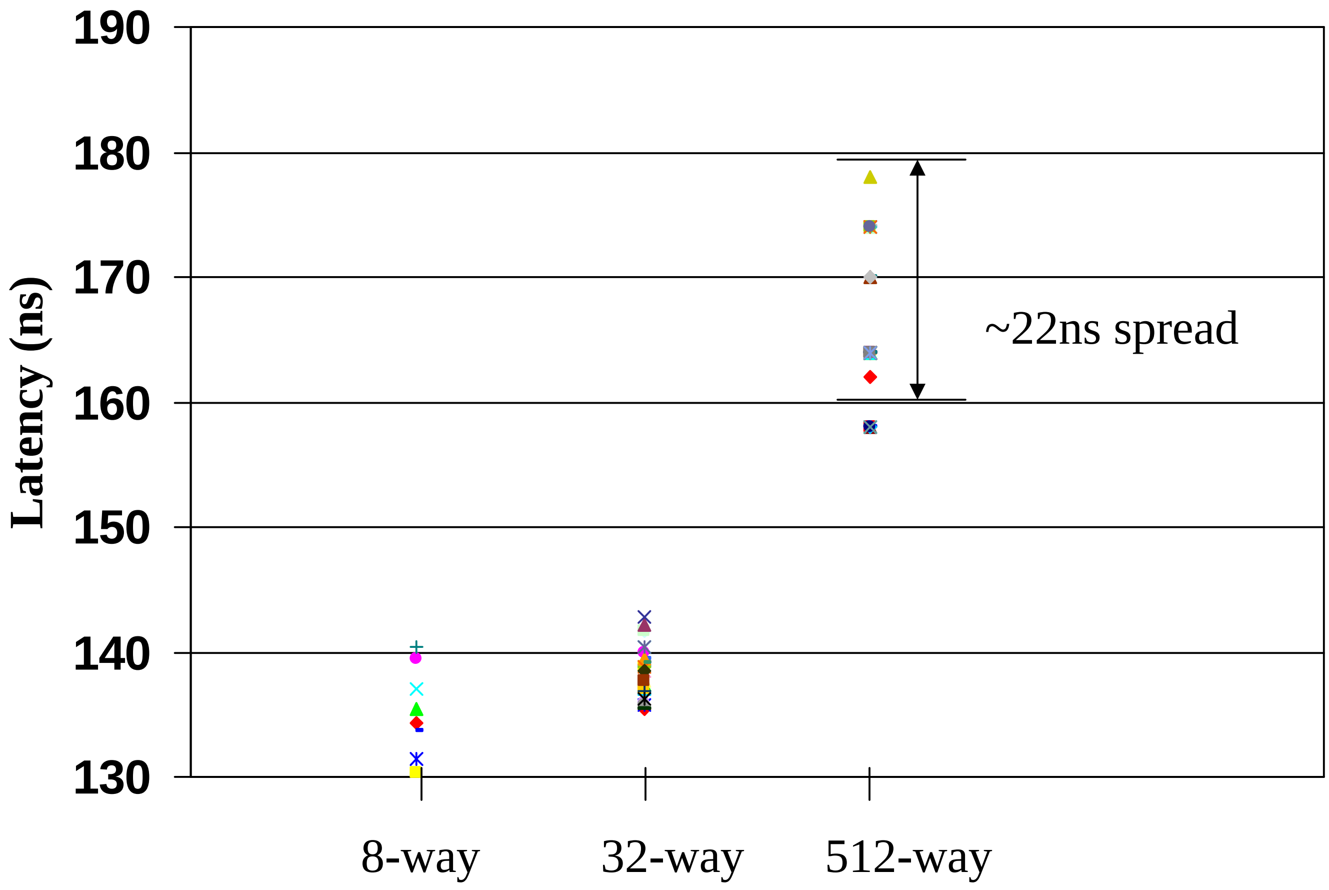


- **Four Independent Barrier or Interrupt Channels**
  - **Independently Configurable as "or" or "and"**
- **Asynchronous Propagation**
  - **Halt operation quickly (current estimate is 1.3usec worst case round trip)**

    **> 3/4 of this delay is time-of-flight.**
- **Sticky bit operation**
  - **Allows global barriers with a single channel.**
- **User Space Accessible**
  - **System selectable**
- **Partitions along same boundaries as Tree, and Torus**
  - **Each user partition contains it's own set of barrier/ interrupt signals**

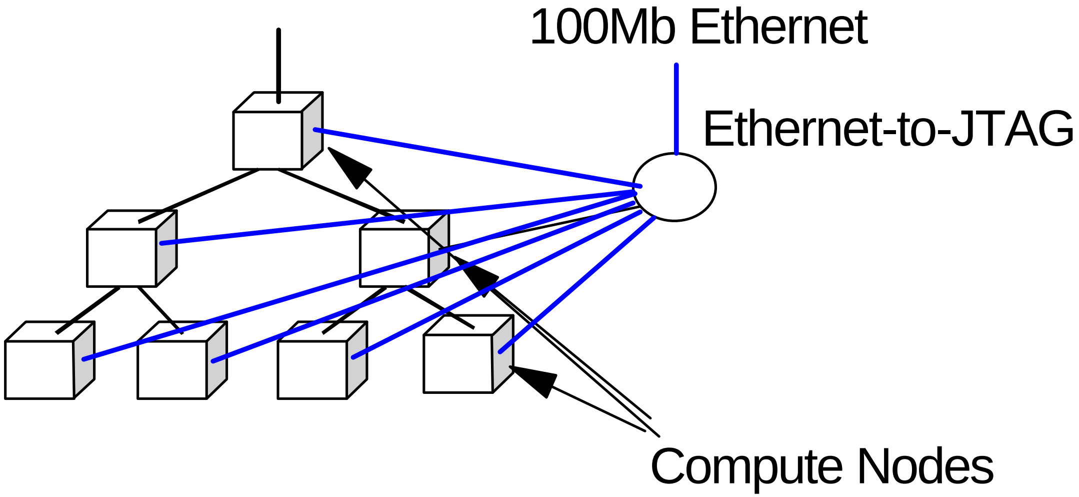
# Global Barriers: Latency and Spread

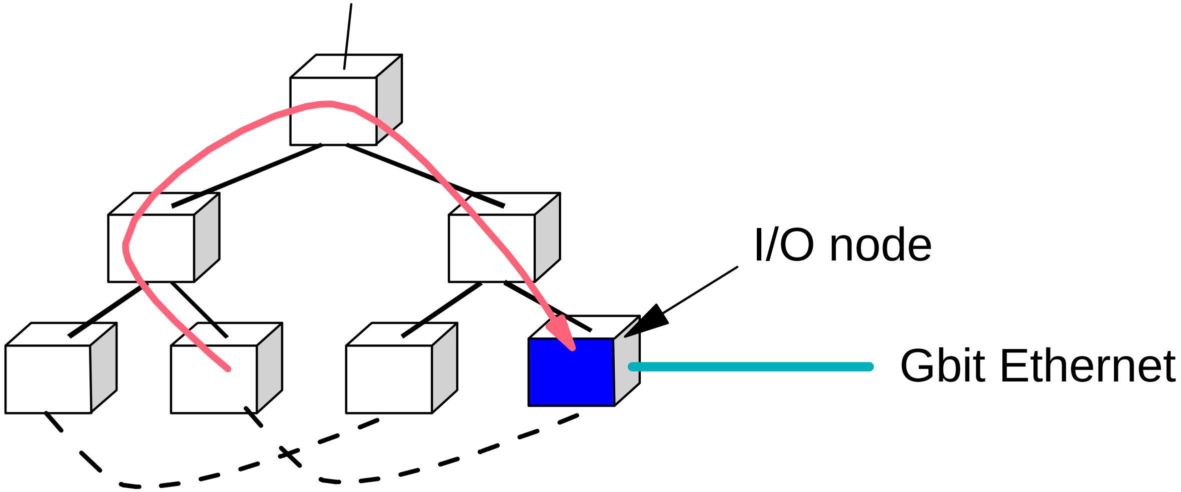
Measured results from prototype confirms ability to scale to 64K with sub-microsecond barrier and interrupt latency



## JTAG interface to 100Mb Ethernet

- direct access to all nodes.
- boot, system debug availability.
- runtime noninvasive RAS support.
- non-invasive access to performance counters
- Direct access to shared SRAM in every node





## Gb Ethernet on all I/O nodes

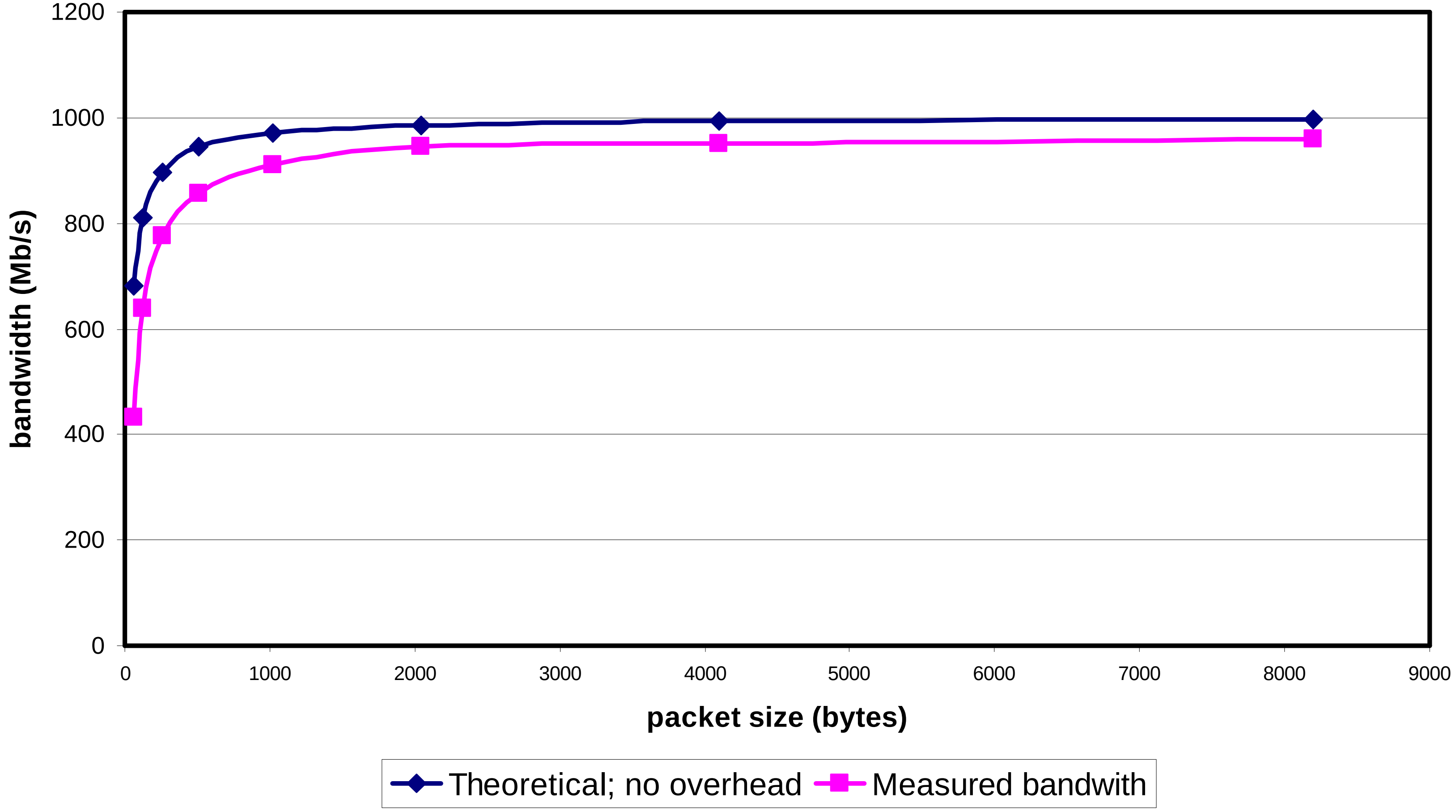
- Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.
- Funnel via global tree.
- I/O nodes use same ASIC but are dedicated to I/O Tasks.
- I/O nodes can utilize larger memory.

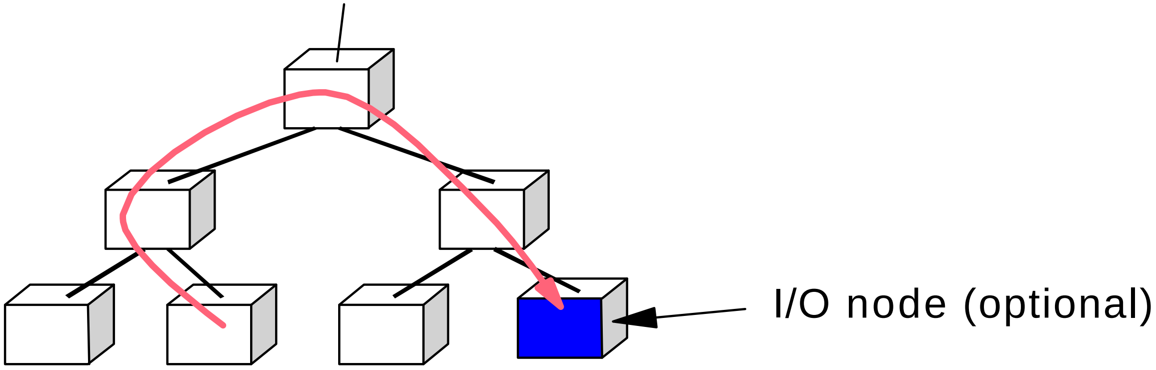
## Dedicated DMA controller for transfer to/from Memory

## Configurable ratio of Compute to I/O nodes

- I/O nodes are leaves on the tree network

# Ethernet Performance is at 97% of Peak



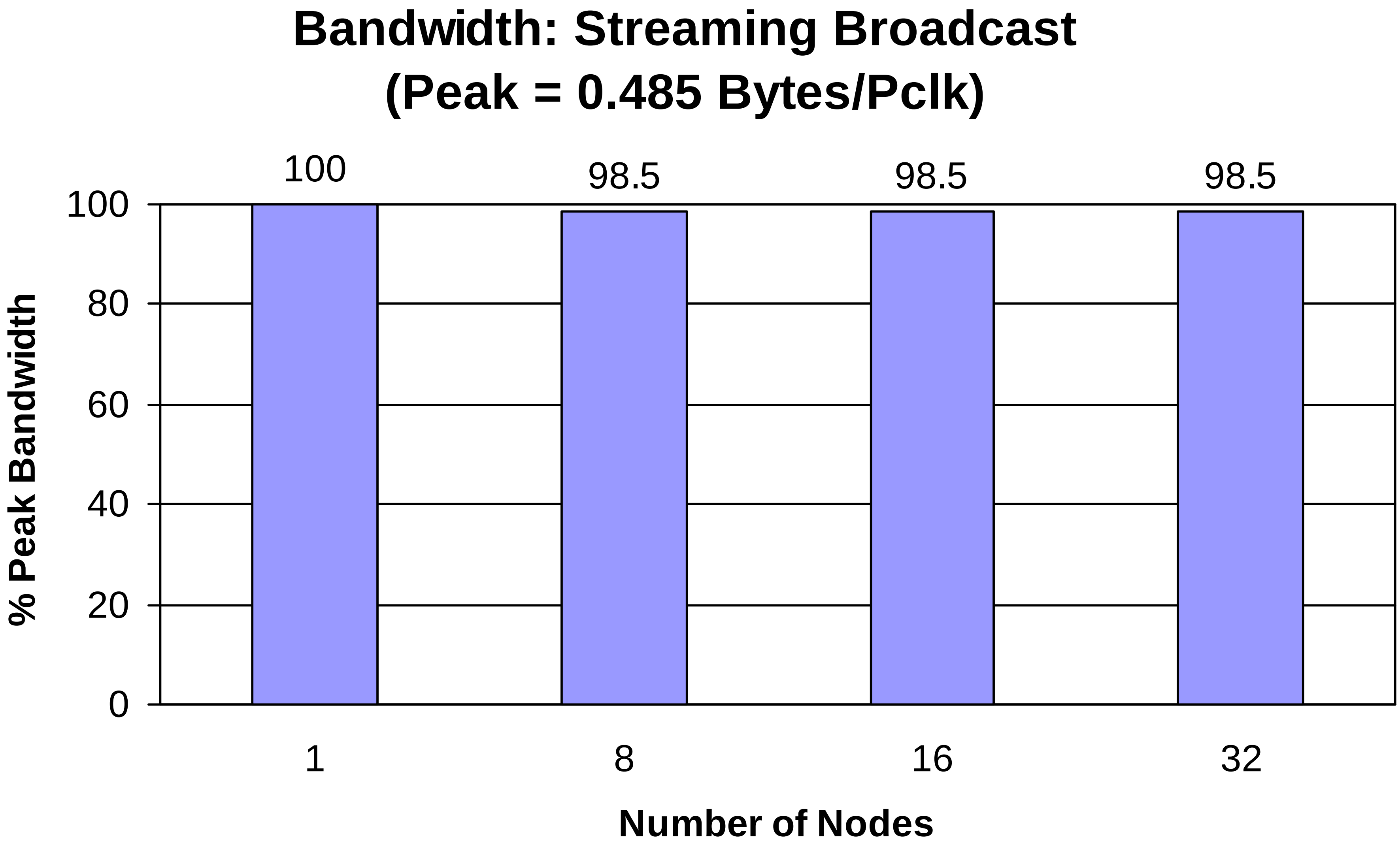


- **High Bandwidth one-to-all**
  - **2.8Gb/s to all 64k nodes**
  - **68TB/s aggregate bandwidth**
- **Arithmetic operations implemented in tree**
  - **Integer/ Floating Point Maximum/Minimum**
  - **Integer addition/subtract, bitwise logical operations**
- **Latency of tree less than 2.5usec to top, additional 2.5usec to broadcast to all**
- **Global sum over 64k in less than 2.5 usec (to top of tree)**
- **Used for disk/host funnel in/out of I/O nodes.**
- **Minimal impact on cabling**
- **Partitioned with Torus boundaries**
- **Flexible local routing table**
- **Used as Point-to-point for File I/O and Host communications**

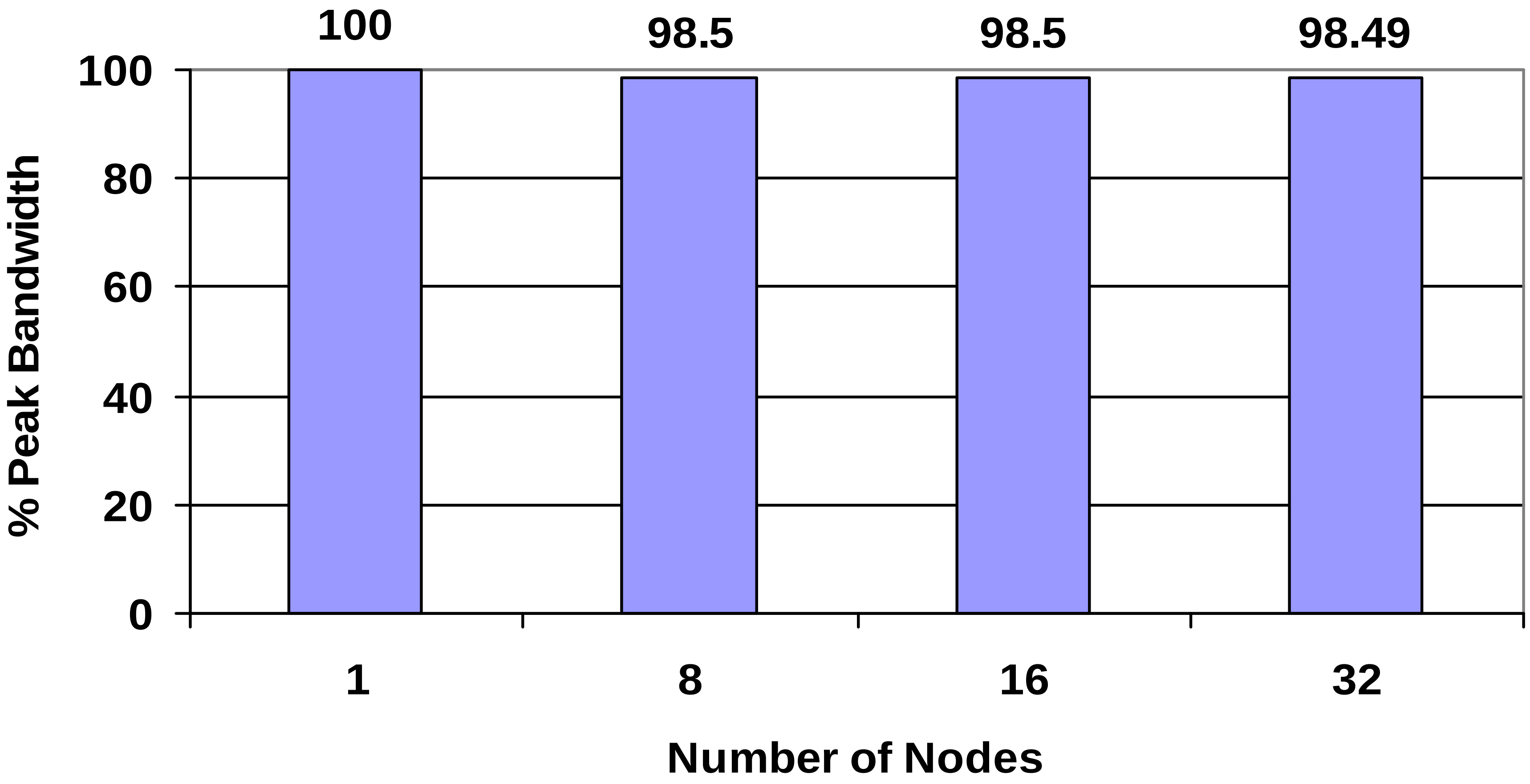
# TREE features cont…

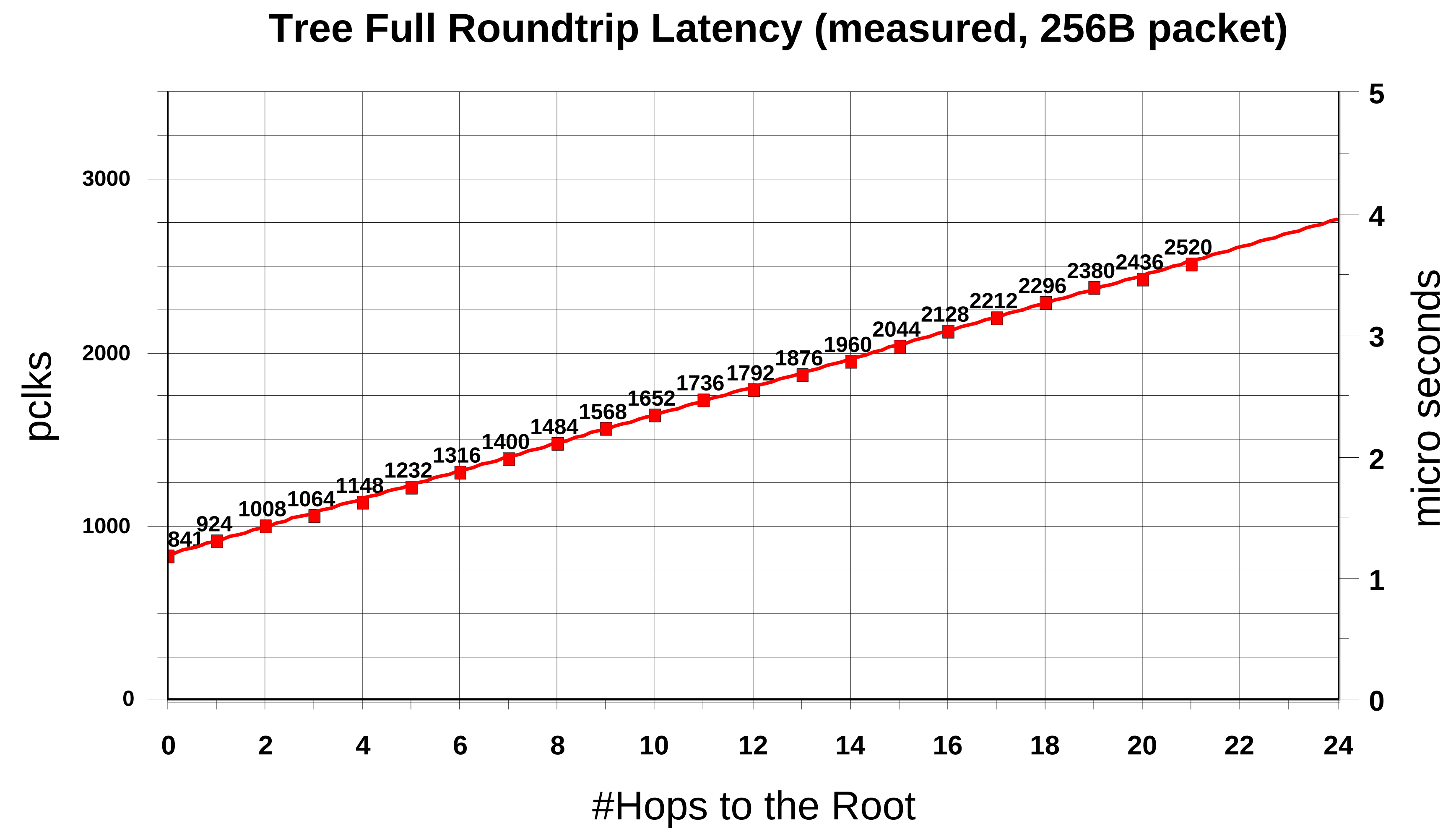
- 350 MB/s per link per direction
  3 network ports -> total: 2.1 GB/s
- Reliable Packet Transport (24bit Packet CRC)
  + 32bit Link CRC
  + Injection/Reception Checksums
- Cut-through routing
  (~60ns latency per hub)
- 256 Bytes payload per packet
- 16 classes / routes per node
- 2 independent virtual networks
- integrated ALU (AND,OR,XOR,MAX,ADD)
  Datatypes: 16bit ... 2048bit (unsigned integer)
- Global combine operation
  64k -> 1 in less than 2.5 us

## Delivered Broadcast bandwidth on Tree Network is consistent with pipelined architecture and should scale to 64K



# Bandwidth: Streaming Reduction (Peak = 0.485 Bytes/Pclk)





R-square = 1 # pts = 17
y = 837 + 80.5x

## Design & Verification:

Dan Beece, Matt Blumrich, Dong Chen,
Marc Dombrowa, Steve Douskey,
Matt Ellavsky, Alan Gara, Jim Goldade,
Mike Hamilton, Ruud Haring, Phil Heidelberger
Dirk Hönicke, Jim Marcella, Ben Nathanson
Martin Ohmacht, Burkhard Steinmacher-Burrow, Sarabjeet Singh, Todd Takken
Brett Tremaine, Mickey Tsao, Pavlos Vranas,
Chuck Wait, Mike Wazlowski
and others …